

# Using Platform-provided Hints for Hot Page Detection and Promotion

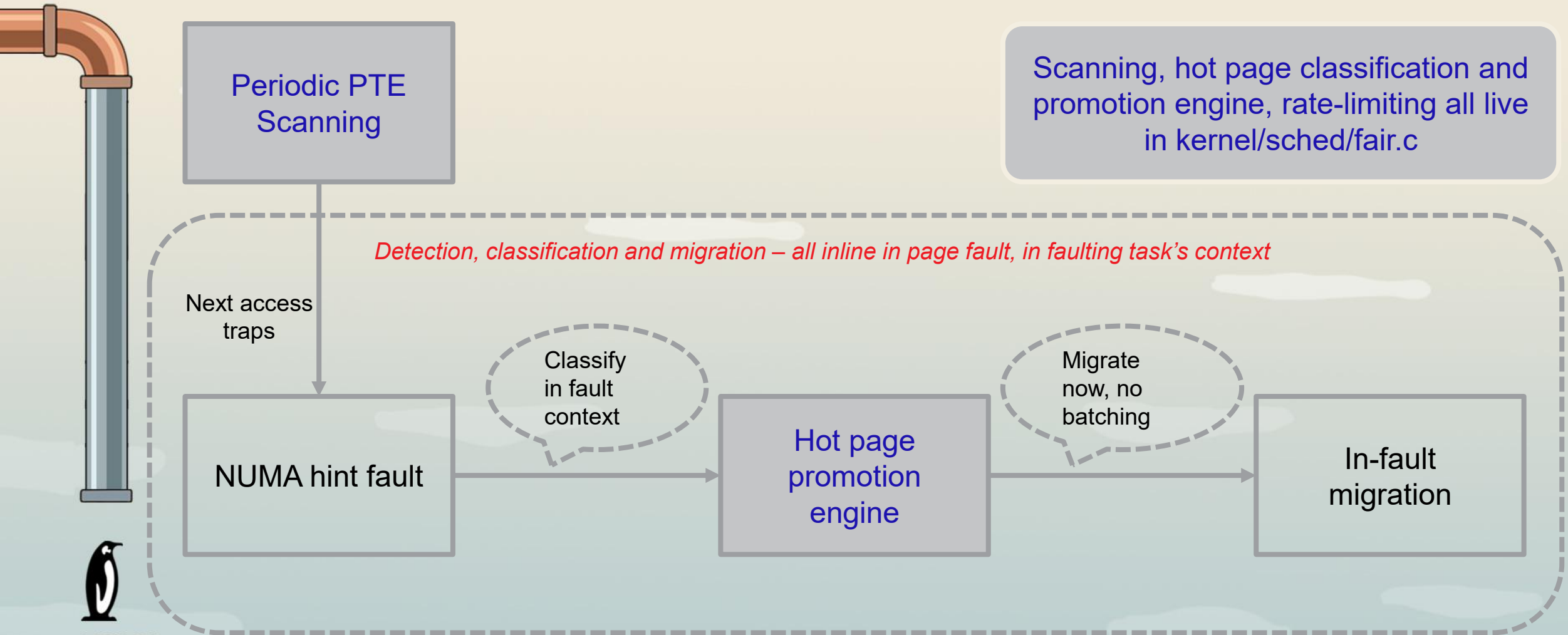
**Device and Specific Purpose Memory MC**  
**Bharata B Rao, AMD**



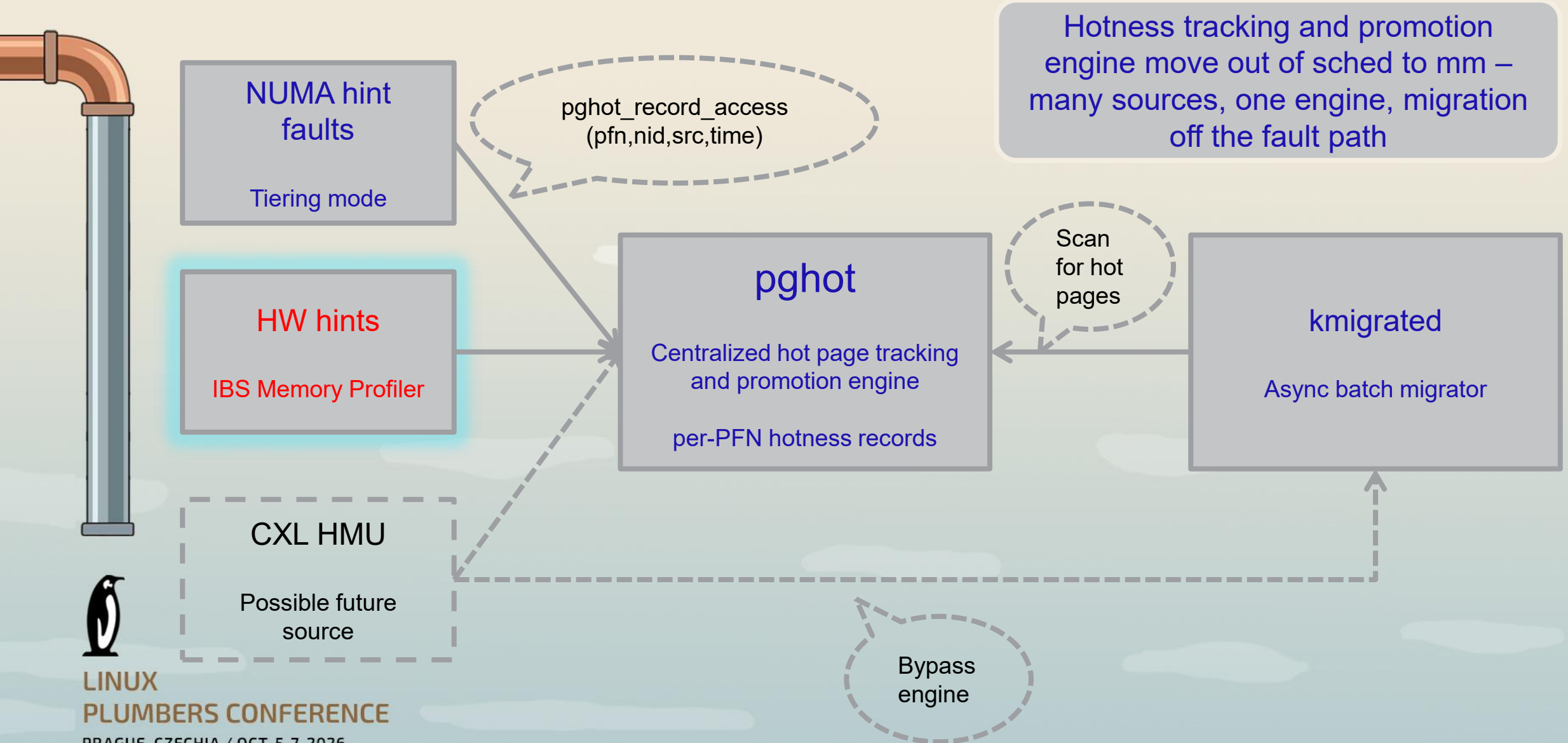
**LINUX  
PLUMBERS  
CONFERENCE**

PRAGUE, CZECHIA / OCT. 5-7, 2026

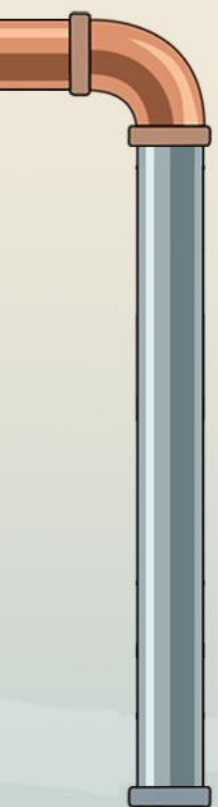
# Hot page promotion – current landscape



# Hot page promotion – possible landscape: pghot



# IBS Memory Profiler – Dedicated memory access sampler

- 
- A separate IBS facility on Zen 6, alongside the existing Fetch and Execution IBS
    - Dedicated for data accesses only
  - Statistical sampling, by op count or cycles
    - Sampling rate is software-tunable, trading coverage against overhead
  - Its own interrupt source
    - Distinct from Fetch and Execution IBS
  - No instruction-fetch accesses, no virtualization support, and no IBS buffering
  - Presence is advertised through a CPUID feature bit

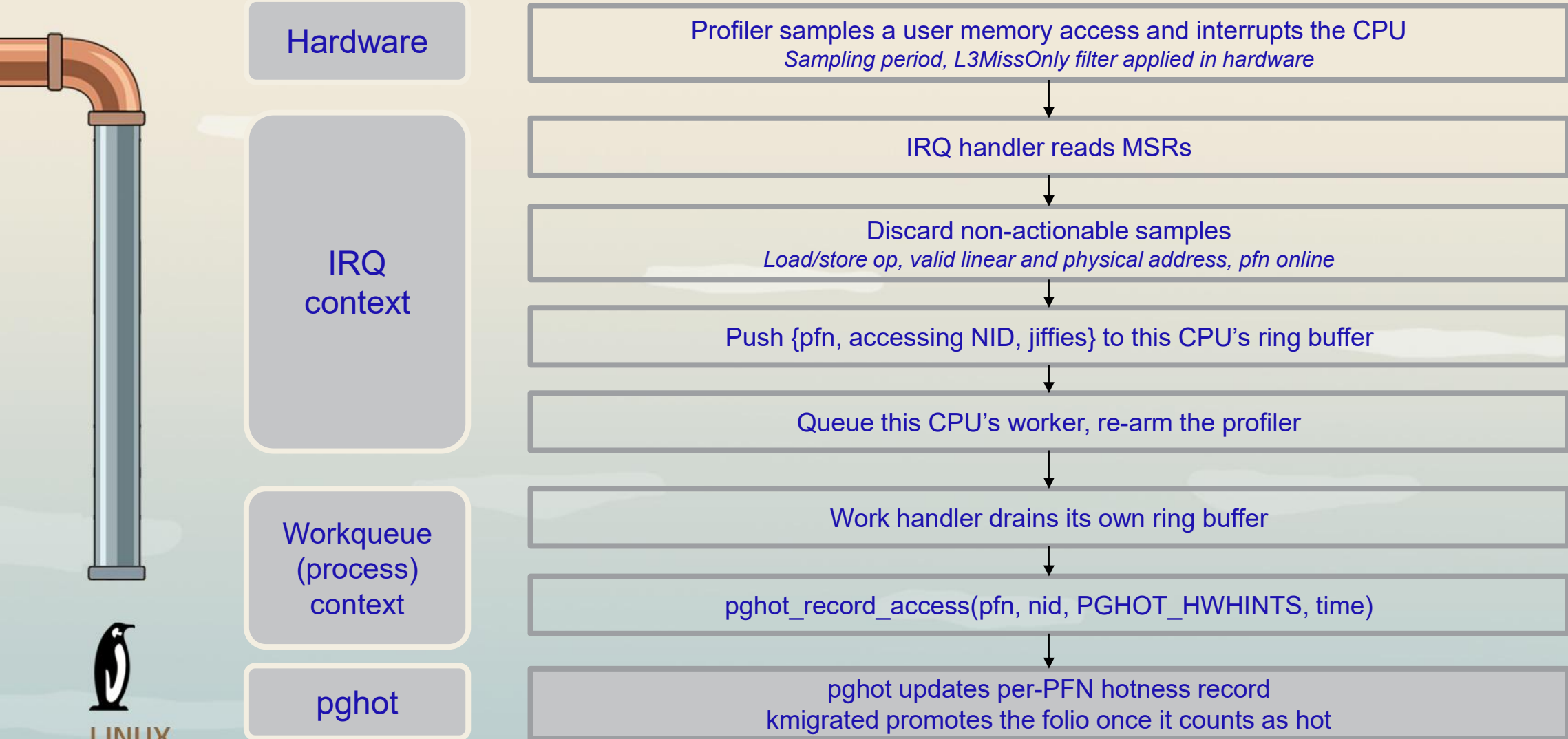


# IBS Memory Profiler – Per-sample data

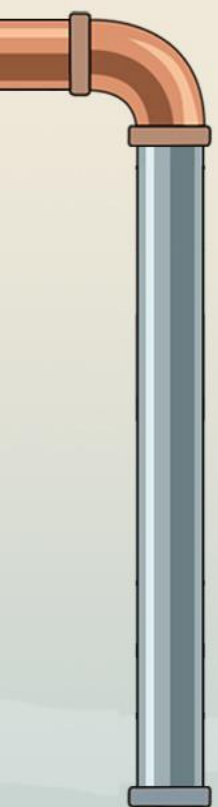
- Each sample identifies the access precisely
  - Linear address of the instruction that issued the access
  - Linear address of the data operand
  - Physical address of the data operand — the field that makes this usable for page hotness
- It also says where the data came from
  - Data cache miss and L2 miss indications
  - Local DRAM, remote DRAM and external memory sources such as CXL are distinguished
- Hardware filtering keeps the interrupt volume manageable
  - L3-miss-only, so only accesses that actually reach memory are reported
  - User-only or kernel-only selection, done by filtering on the instruction pointer
  - Load latency filtering
- Why this fits tiered memory?
  - The kernel is handed a physical address, without scanning page tables, without inducing faults



# IBS Memory Profiler – Sample flow



# IBS Memory Profiler numbers: Non-overcommitted scenario

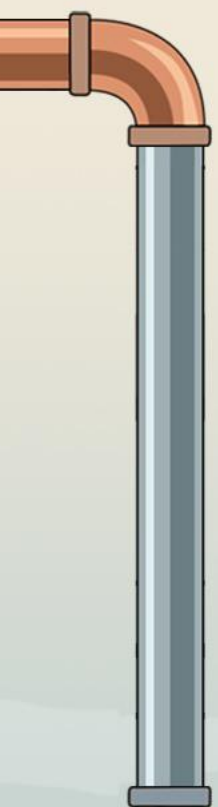


| Benchmark       | Metric (higher is better) | base-NUMAB2 | pghot-hwhints (IBS Memory Profiler) |
|-----------------|---------------------------|-------------|-------------------------------------|
| NAS BT (MPI)    | Mop/s total               | 2.34x       | 2.15x (Near parity)                 |
| Graph500 BFS    | harmonic-mean TEPS        | 2.34x       | 3.17x (Win)                         |
| Redis + memtier | ops/sec                   | 1.04x       | 1.00x (Miss)                        |
| XSbench         | Lookups/s                 | 1.18x       | 4.57x (Win)                         |
| Pointer-chase   | 1/latency (p10k/p5k)      | 2.24x       | 1.49x/3.48x (Miss/Win)              |

- All results are on 3-node tiered system (DRAM top tier + CPU-less CXL lower tier)
  - 256CPUs, 792GB RAM, 256GB on CXL
- Speedups below are normalized to the base kernel with no tiering (NUMAB=0)
- Columns compare mainline hint-fault tiering (base NUMAB=2) against IBS memory profiler (pghot-hwhints, NUMAB=0, no NUMA scanning)
- The working set is moved to CXL tier before the run phase, top tier has headroom, so **no demotion**



# IBS Memory Profiler numbers: Overcommitted scenario



| Benchmark    | Metric (higher is better) | base-NUMAB2 | pghot-hwhints (IBS Memory Profiler) |
|--------------|---------------------------|-------------|-------------------------------------|
| Graph500 BFS | harmonic-mean TEPS        | 1.04x       | 1.07x (Near parity)                 |
| llama.cpp    | decode tok/s (tg128)      | 1.19x       | 1.22x (Near parity)                 |
| XSbench      | Lookups/s                 | 0.98x       | 0.96x (Near parity)                 |

- All results are on 3-node tiered system (DRAM top tier + CPU-less CXL lower tier)
  - 256CPUs, 792GB RAM, 256GB on CXL
- Speedups below are normalized to the base kernel with no tiering (NUMAB=0)
- Columns compare mainline hint-fault tiering (base NUMAB=2) against IBS memory profiler (pghot-hwhints, NUMAB=0, no NUMA scanning)
- The working set spills over to lower tier and hence **both promotion and demotion** are involved



# IBS Memory Profiler – Overhead measurements

Micro-benchmark, 64 threads, 8G on CXL, random repetitive accesses resulting in promotion. Benchmark score – completion time, lower is better

| Metric              | base-NUMAB2 | pghot-hwhints |
|---------------------|-------------|---------------|
| Benchmark score (s) | 37.853      | 39.906        |
| elapsed_s           | 69.76       | 69.47         |
| user_s              | 2267.46     | 2262.30       |
| system_s            | 64.96       | 6.74 (*)      |
| pgpromote_success   | 2,097,152   | 1,928,510     |

(\*) Excludes IRQ and kmigrated CPU time



LINUX  
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

# IBS Memory Profiler – Overhead measurements

base-NUMAB2

| Path                                | CPU-s  | Events                |
|-------------------------------------|--------|-----------------------|
| numa_scan_ns                        | 0.160  | 2,097,152 PTE updates |
| numa_hint_fault_ns                  | 26.355 | 2,427,770 hint faults |
| numa_migrate_ns                     | 28.203 | 2,097,152 pages       |
| sum                                 | 54.718 |                       |
| Normalized per-folio promotion cost | 26.1us |                       |

pghot-hwhints

| Path                                | CPU-s  | Events               |
|-------------------------------------|--------|----------------------|
| ibs_mprof_irq_ns                    | 7.176  | 5,081,598 IBS IRQs   |
| pghot_record_access_ns              | 3.698  | 5,081,598 reported   |
| kmigrated_pfn_walk_ns               | 14.248 | 1,932,182 recorded   |
| kmigrated_batch_migrate_ns          | 5.257  | 1,928,510 promotions |
| sum                                 | 30.379 |                      |
| Normalized per-folio promotion cost | 15.8us |                      |



LINUX  
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

CPU-s: CPU nanoseconds summed up, converted to seconds

# Discussion points

- What should be common?

- Hotness interpretation, or only async migration/rate-limiting?
- Should pghot be one subsystem or an optional hotness engine over kmigrated?

- Do all sources fit one model?

- Raw samples: hint faults, IBS
- Ranked pages/ranges: CHMU
- One API or separate sample and ready-candidate APIs?
- Page vs region observations: DAMON

- Other open questions

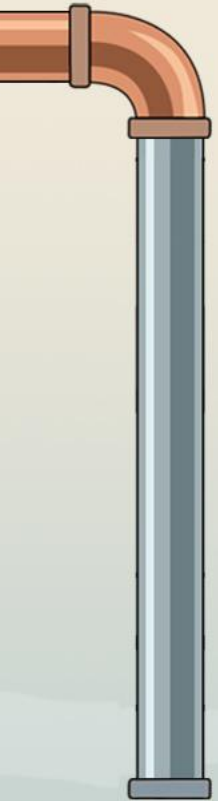
- Do sources need to coexist or only share migration backend?
- If sources coexist, how are unequal or duplicate observations combined?
- Must ranked sources pay for per-PFN state?
- Under pressure, must promotion be coordinated with demotion?

- What is the acceptance bar?

- NUMAB2 move: parity
- New source: workload win or overhead reduction
- Shared engine: coexistence/reuse benefit
- Async migrator: reduce CPU/PSI/tail latency without throughput regression?



# What should survive as the common abstraction?



- Full pghot
  - Shared hotness engine and shared migrator
- Split model
  - Optional hotness engine over a common kmigrated backend
  - Separate APIs for sampled and ranked sources
- Migrator only
  - Sources classify hotness independently and submit pages/ranges
- No new subsystem
  - Extend NUMAB2, DAMON and hardware-source paths independently



LINUX  
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

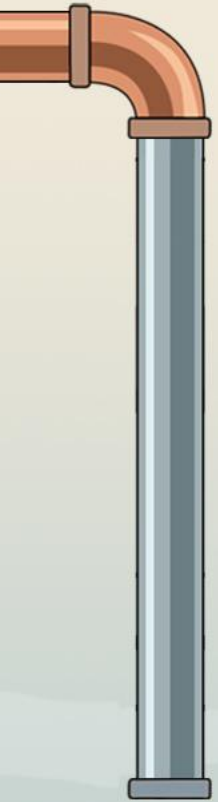
**Working hypothesis:** A common kmigrated backend, with an optional hotness engine and separate APIs for sampled and ranked sources

# Backup



**LINUX**  
**PLUMBERS CONFERENCE**  
PRAGUE, CZECHIA / OCT. 5-7, 2026

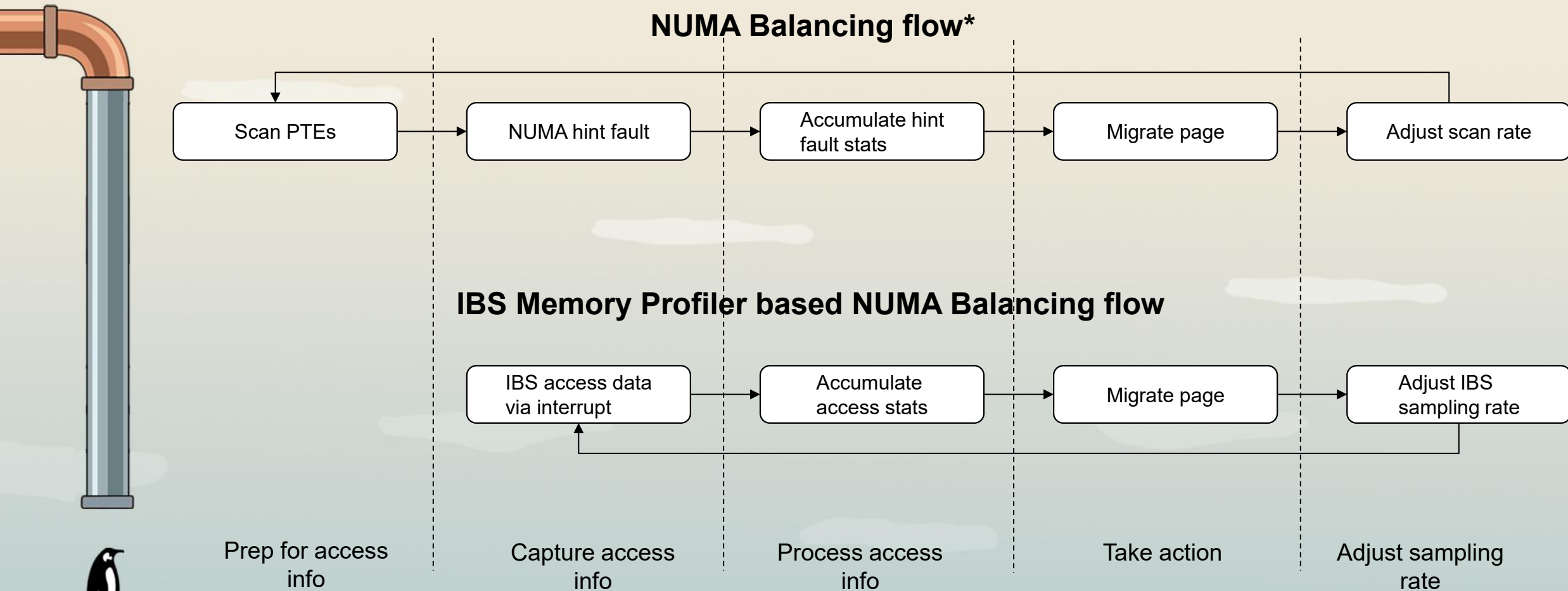
# IBS Memory Profiler – Dedicated memory access sampler



- Same programming model as primary IBS, but its own control and enable
  - Light weight: Keeps only what is relevant to data memory operations
- Sampling
  - HW tags one op out of every N dispatched ops, or every N core clock cycles, and profiles it
  - When the tagged op retires and passes the enabled filters, the sample is latched and an interrupt is raised
  - SW reads the sample and re-arms the profiler, which starts the next interval



# Alternative: IBS Memory Profiler feeding NUMA hint faults



Prep for access info

\* Simplified view

Capture access info

Process access info

Take action

Adjust sampling rate



LINUX  
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

# NAS BT – Non overcommitted

NAS Parallel Benchmark BT (Block Tri-diagonal solver), MPI version, Class D, 16 ranks  
Provisioned memory (after migration)  $\approx$  161.378 GiB

| Metric (in millions)    | base-NUMAB2 | pghot-hwhints |
|-------------------------|-------------|---------------|
| pgpromote_success       | 42.18       | 39.70         |
| pgpromote_candidate_nrl | 42.18       | 39.70         |
| numa_pte_updates        | 42.25       | -             |
| numa_hint_faults        | 42.18       | -             |
| pghot_recorded_accesses | -           | 39.79         |
| pghot_reported_hwhints  | -           | 553.94        |
| hwhint_dram_accesses    | -           | 513.14        |
| hwhint_extmem_accesses  | -           | 39.79         |



# Graph500 – Non overcommitted

Graph500 reference BFS, SCALE=28, edgefactor=16, 128 ranks  
Provisioned memory: ~100GB

| Metric (in millions)    | base-NUMAB2 | pghot-hwhints |
|-------------------------|-------------|---------------|
| pgpromote_success       | 13.32       | 3.70          |
| numa_pte_updates        | 25.87       | -             |
| numa_hint_faults        | 13.32       | -             |
| pghot_recorded_accesses | -           | 3.71          |
| pghot_reported_hwhints  | -           | 21.38         |
| hwhint_dram_accesses    | -           | 16.79         |
| hwhint_extmem_accesses  | -           | 3.70          |



LINUX  
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

# Graph500 – Overcommitted

Graph500 reference BFS, SCALE=30, edgefactor=16, 128 ranks  
Provisioned memory: ~400GB

| Metric (in millions) | base-NUMAB2 | pghot-hwhints |
|----------------------|-------------|---------------|
| pgpromote_success    | 32.02       | 6.75          |
| pgdemote_kswapd      | 32.75       | 9.79          |



LINUX  
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

# Redis + Memtier – Non overcommitted

~64GB data set (62.2M keys x 1kB)

GET traffic with memtier: 16 threads x 100 conns, PASSES=12 over 50% of keyspace

| Metric (in millions)    | base-NUMAB2 | pghot-hwhints |
|-------------------------|-------------|---------------|
| pgpromote_success       | 10.43       | 0.57          |
| numa_pte_updates        | 20.33       | -             |
| numa_hint_faults        | 10.43       | -             |
| pghot_recorded_accesses | -           | 1.11          |
| pghot_reported_hwhints  | -           | 1.97          |
| hwhint_dram_accesses    | -           | 0.85          |
| hwhint_extmem_accesses  | -           | 1.11          |



# Pointer chase – Non overcommitted

64-thread pointer chase, 64 GB total buffer, 16 GB hot set, 240 s timed phase  
Buffer relocated to node 2 before the timed phase

| Metric (in millions)    | base-NUMAB2 | pghot-hwhints |
|-------------------------|-------------|---------------|
| pgpromote_success       | 3.79        | 1.29          |
| numa_pte_updates        | 12.91       | -             |
| numa_hint_faults        | 3.79        | -             |
| pghot_recorded_accesses | -           | 1.29          |
| pghot_reported_hwhints  | -           | 1.40          |
| hwhint_dram_accesses    | -           | 0.11          |
| hwhint_extmem_accesses  | -           | 1.29          |

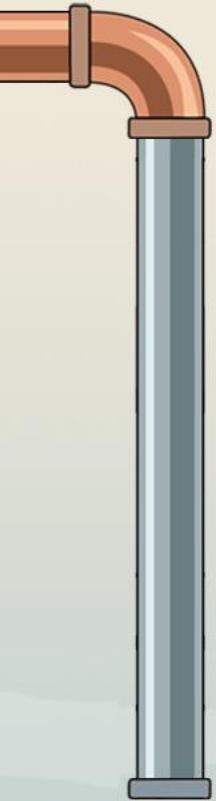


# Llama.cpp – Overcommitted

llama-bench (llama.cpp), Mixtral-8x22B-Instruct Q4\_K\_M (140.6B params, 79.7 GB), -t 64 -p 512 -n 128 -r 5 --mmap 0.

A 200 GB hot hog on toptier forces kswapd to naturally demote ~1/3 of the model to CXL

| Metric (in millions)    | base-NUMAB2 | pghot-hwhints |
|-------------------------|-------------|---------------|
| pgpromote_success       | 3.07        | 1.74          |
| pgpromote_candidate     | 12.80       | 12.86         |
| pgdemote_kswapd         | 4.78        | 4.40          |
| numa_pte_updates        | 26.06       | -             |
| numa_hint_faults        | 25.12       | -             |
| pghot_recorded_accesses | -           | 12.86         |
| pghot_reported_hwhints  | -           | 30.79         |
| hwhint_dram_accesses    | -           | 17.74         |
| hwhint_extmem_accesses  | -           | 12.86         |



# XSbench – Non overcommitted

XSbench: -s XL -g 238847 -t 64 -p 60000000 -l 34 (history, unionized grid), footprint  
~116.56 GiB

| Metric (in millions)    | base-NUMAB2 | pghot-hwhints |
|-------------------------|-------------|---------------|
| pgpromote_success       | 30.53       | 2.70          |
| numa_pte_updates        | 30.56       | -             |
| numa_hint_faults        | 30.54       | -             |
| pghot_recorded_accesses | -           | 2.71          |
| pghot_reported_hwhints  | -           | 88.44         |
| hwhint_dram_accesses    | -           | 84.52         |
| hwhint_extmem_accesses  | -           | 2.68          |



LINUX  
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

# XSbench – Overcommitted

XSbench: -s XL -g 615000 -t 64 -p 60000000 -l 34 (history, unionized grid), footprint  
~300.13 GiB

| Metric (in thousands)   | base-NUMAB2 | pghot-hwhints |
|-------------------------|-------------|---------------|
| pgpromote_success       | 0           | 244.98        |
| pgpromote_candidate     | 1           | 267.64        |
| pgdemote_kswapd         | 0           | 304.27        |
| numa_pte_updates        | 1           | -             |
| numa_hint_faults        | 1           | -             |
| pghot_recorded_accesses | -           | 267.65        |
| pghot_reported_hwhints  | -           | 89870         |
| hwhint_dram_accesses    | -           | 88940         |
| hwhint_extmem_accesses  | -           | 272.63        |



# IBS Memory Profiler – THP Results summary, Non overcommitted case

| Benchmark     | Metric (higher is better) | base-NUMAB2 | pghot-hwhints (IBS Memory Profiler) |
|---------------|---------------------------|-------------|-------------------------------------|
| NAS BT (MPI)  | Mop/s total               | 2.38x       | 2.33x                               |
| Pointer-chase | 1/latency (p10k)          | 2.31x       | 1.46x                               |



LINUX  
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

# References

- IBS Memory Profiler – [specification document](#)
- pghot + IBS Memory Profiler [patchset](#)
- IBS Memory Profiler RFC v0 [patchset](#)



LINUX  
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

# COPYRIGHT AND DISCLAIMER

©2026 Advanced Micro Devices, Inc. All rights reserved.

AMD, the AMD Arrow logo and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate releases, for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

THIS INFORMATION IS PROVIDED 'AS IS.' AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.



**LINUX**  
**PLUMBERS CONFERENCE**

PRAGUE, CZECHIA / OCT. 5-7, 2026